**PH1976 Fundamentals of Data Analytics and Predictions**

**Spring 2023 - Class Project**

**Predicting Parkinson’s Disease for Patients Using Voice Recording**

The course project can be done in groups of 3-5 students. Each group should be diverse with respect to the student’s major. Therefore, each group should have at least one non-Biostatistics major student.

The project will consist of an in-class oral presentation by two or more of the group members. The day of the presentation, the instructor will randomly choose which of the members will present the project so all members should be prepared to present.

1. Group or individual 15 minute (10 minutes for presentation, 5 minutes for questions) in-class oral presentations will be scheduled on **April 17, 19, and 26, 2023**. **Note no class on April 24 due to a conflict.**
2. The due date for your report, code, and predictions is by midnight on **April 15, 2022**, regardless of when your presentation is scheduled.

* Oral presentation. 15 minutes (10 minutes for presentation, 5 minutes for questions).
  + All
* Detailed report of your work with introduction, methods, results (with illustrations), and discussions with a maximum of 5 pages.
  + Code:
  + Introduction:
  + Methods:
  + Results:
  + Discussions:
* Program source code
  + Make sure to include enough documentations and comments in your code. An R Markdown file is preferable.
* Test set predictions, an R dataset that has one column with the patient ID numbers and the other column with their predicted outcome of mortality.

The project will be graded as

* 10% internal group evaluation,
* 40% Written report as graded by the professors,
* 30% oral presentation as graded by the professors,
* 15% oral presentation as graded by fellow students,
* 5% (up to 5pts bonus) based on algorithm performance on the left-out test set based on the ranking among other groups (e.g., best team takes 5 points + 5 bonus points (10/5) and the worst performing team takes (0/5)).

Every student will fill out an evaluation form of every other presentation as well as an internal evaluation of your group members. Attendance of other group’s presentations is mandatory. Evaluations will be anonymous to other students. If a student does not submit attendance, that student’s final project grade will be deducted by 3 points.

Once you form your group, send all names in the group to the TAs who will assign a group number.

The talk will be split up into a background/methods section, results section, and a discussion section, each presented by a randomly chosen person from the group (so that everyone is prepared to present all sections, we encourage you to practice in front of each other).

The goal of this project is to predict Parkinson’s disease (PD) using the extracted features from the voice recording of patients. For each individual, three recording samples were collected, students can perform data pre-processing steps before making the final prediction on a patient.A test set will be given but the outcome will not be included in the dataset.

**Definition**

Parkinson’s disease (PD) is a progressive neurodegenerative disorder. To accurately detect the disease in the early stage, many telediagnosis and telemonitoring systems have recently been proposed. Since vocal problem is one of the most important symptoms which can be seen in the earlier stage of PD patients, vocal disorders-based systems become popular in PD diagnosis and monitoring. In these systems, various speech signal processing algorithms have been used to extract clinically useful information for PD assessment, and the calculated features are fed to different learning algorithms to make reliable decisions. PD telemedicine studies showed that the choice of extracted features and learning algorithms directly influences the accuracy and reliability of distinguishing PD patients.

**Data Source**

In this study, we collected the voice recordings of 252 subjects including PD patients and healthy individuals. We got three recording samples from each subject and extracted seven feature subsets from the recording samples. The feature subsets were baseline features, intensity-based features, bandwidth and formant features, vocal fold features, Mel Frequency Cepstral Coefficients (MFCC), wavelet transform based features (WT) and tunable Q-factor wavelet transform based features (TQWT).

**Study Design**

The features subsets are fed to multiple classifiers including SVMs with linear and RBF kernels, Multilayer Perceptron, Naive Bayes, Logistic Regression, Random Forest and k-NN algorithms. We separate the dataset into a training set (n = 176) and a test set (n = 76).

**Study Population**

The dataset includes PD patients with age ranging from 33 to 87 (65.1 ± 10.9) and healthy individuals with age ranging from 41 to 82 (61.1 ± 8.9). Each patient has three voice recording samples, with 7 aforementioned feature subsets. Each feature subset contains several features.

**Outcome:**

The outcome is the variable ‘class’: 0 means without PD and 1 means with PD.

**Task:**

Build a machine-learning model to predict PD based on the voice recordings of patients.

**Data:**

A link will be sent out on Canvas.

Github notes: PH 1976 Repository, DEVELOPMENT branch. Git GUI is an app for your computer that allows you to ‘commit’ changes to code directly to Github once linked! 😊

Sara: start on report document (methods? Did I get this right?) and R markdown

Erin: R Markdown file started. Rough report outline assembled, taken from assignment instructions. Will need updating by the respective parties. Initial coding complete: file loading, package installation, split into training/testing data, pre-processing/standardization, predictor selection, baseline logistic regression, and random forest methods. My portion is complete for now! 😊 Will pick back up during report writing, mostly methods.

Safa: Introduction, help with PPT and scripts

Lakshmi: PPT, help with any code

Jackie: 2nd part (new methods) of code; Results section; Discussion

Next Meeting:

**Friday April 7 6:30 pm**

for help- contact gm

We might reconvene to finalize results and discussion together